

CLaC Lab - Sabine Bergler

- Topic: **Natural language processing**
 - Enhance current machine learning models
 - to analyze written text
 - for classification or knowledge extraction
- Opportunities:
 - 1-2 summer USRAs
 - 1-2 Master's or PhD students

Information Extraction

- Named entity extraction in Spanish clinical reports: extract, distinguish, and classify *humans* and *species*

ANTECEDENTES FAMILIARES Padre ingresado desde hace 15 días por sospecha de tuberculosis pulmonar (clínica respiratoria de tos y febrícula de 6 meses de evolución). Afectación en radiografía de tórax y TAC pulmonar y Mantoux positivo. Ha iniciado tratamiento antituberculoso desde hace 10 días. Mantoux a contactos familiares pendientes de leer.

- Adverse drug reaction detection in tweets

saphris gives me a mad appetite omg i hate this

- Multi-label classification of abstracts from Covid-related articles into 7 classes: Treatment, Mechanism, Prevention, Case Report, Diagnosis, Transmission, and Epidemic Forecasting

... to undergo Continuous Positive Airway Pressure (CPAP) or Non-Invasive Positive Pressure Ventilation (NIPPV) due to ...

Sentiment and figurative language detection

- Classify tweets into *positive, negative, or neutral* sentiment carriers

El Classico on a Sunday Night isn't perfect for the Monday Morning !!

- Classify tweets that use non-literal language into positive, negative, or neutral sentiment carriers

A paperless office has about as much chance as a paperless bathroom

- Classify user profiles of 400 tweets into users of irony or stereotype spreaders

Recent contribution

DAM (Parsa Bagherzadeh)

- an architecture
- to combine ML models
- with existing ontologies

- showed parameter space reduction
- solid performance
- visualization possibilities

Current work: Summarization

- Summarization using BART
- DATA: AnsSum shared task: <https://www.aclweb.org/portal/content/shared-task-perspective-aware-healthcare-answer-summarisation-cl4health-workshop-naacl-2025>
- TOOL: BART: <https://www.digitalocean.com/community/tutorials/bart-model-for-text-summarization-part1>

Current work: Information Extraction

- Gut Brain interaction IE shared task at CLEF
- TOOL:
 - BERT <https://machinelearningmastery.com/a-complete-introduction-to-using-bert-models/>
 - T5 https://medium.com/@gagangupta_82781/understanding-the-t5-model-a-comprehensive-guide-b4d5c02c234b
- Data pre-processing
- DATA:
 - GutBrainIE: <https://clef2025.clef-initiative.eu/index.php?page=Pages/Labs/BioASQ.html>

Current work: Classification with Disagreements

- Task: A classification task is annotated by 6 annotators. Demographic features are provided with labels
- Research question: Can the metainformation improve performance?

Future work:

- How can we combine known linguistic features with PLMs to improve classification?
- How can we use facts known about the world to enhance PLMs to improve classification?
- How can we use knowledge to improve prompting?
- How can we prevent catastrophic forgetting when incrementally fine-tuning?
- What is the future of state-enabled LLMs? (Mamba?)